Nature vs. Nature

Can classification models tell the difference?

Discerning between two similarly themed subreddits

By Jack Roach

r/NatureIsMetal

A Cardinal and a Blue Jay have a little confrontation

Posted by u/roastedtoperfection

r/NatureIsFuckingLit

🔥 Shizuoka, Japan

🔥Posted by u/Kris19275

Problem Statement

- Use Natural Language Processing to build a subreddit classification model
- Predict which of these two subreddits a post comes from based on its title
- 5,000 recent posts from each subreddit used to train the model

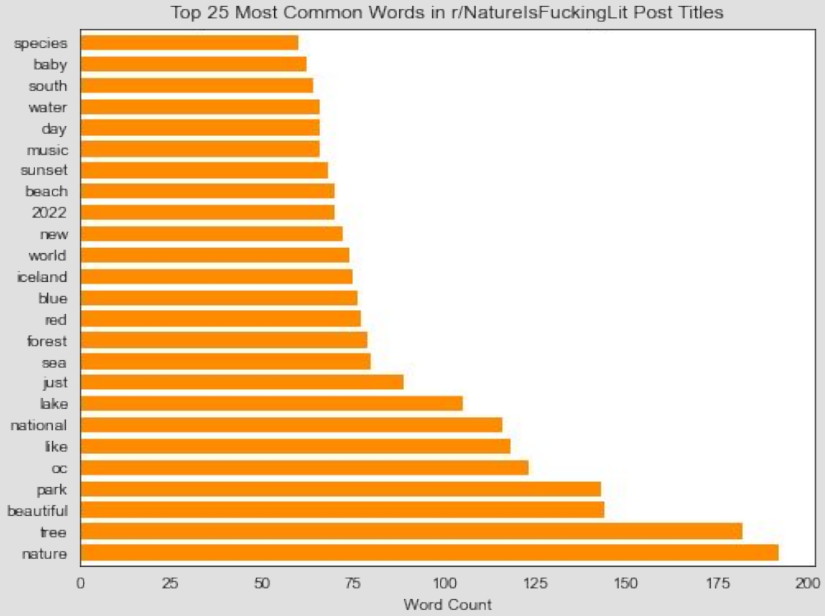
Commonly occurring nature words

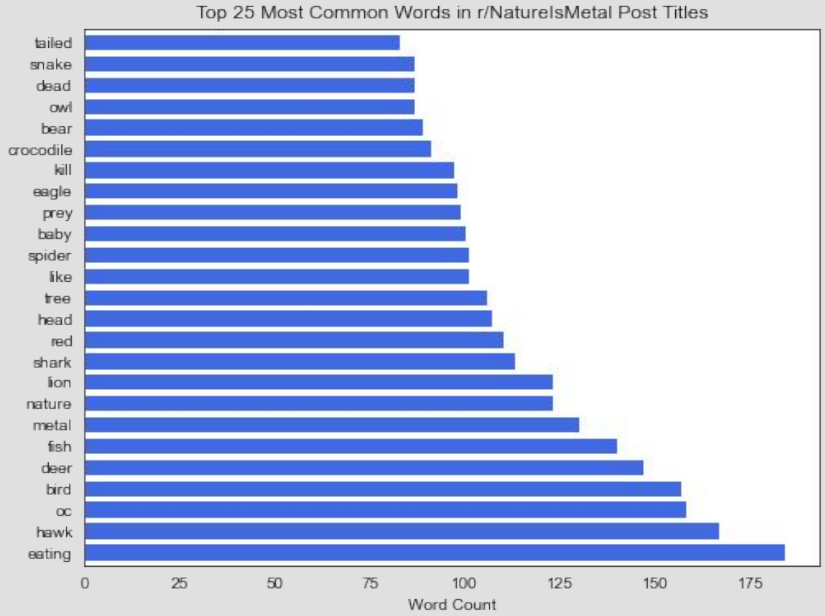
		Count
1.	🔥	3,367
2.	Nature	315
3.	Tree	288
4.	Oc	281
5.	Like	219
6.	Eating	217
7.	Bird	209
8.	Park	207
9.	Hawk	198
10.	Fish	189

r/NatureIsFuckingLit
Rule #2:
Titles must start with the 🔥 emoji

Word Frequency by Individual Subreddit

🔥 excluded for plot clarity



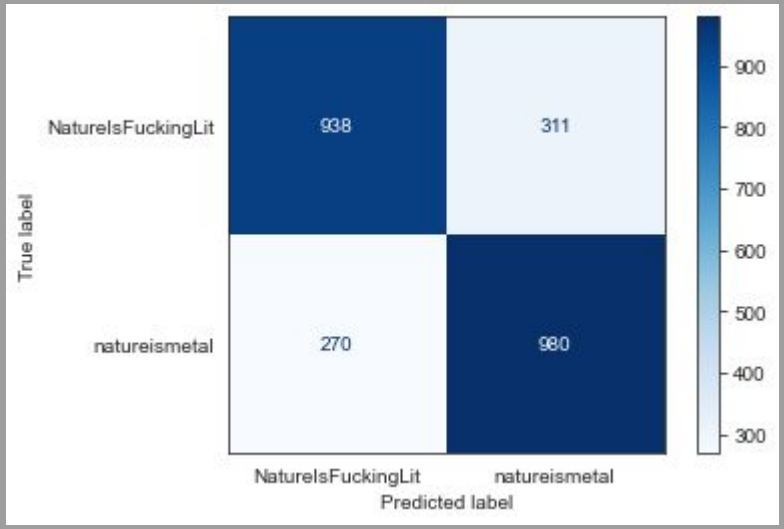


Voting Classifier

Our Model = 76.75%

Evenly weighted models

- Multinomial Naive Bayes with Count Vectorizer = 76.43%
- Logistic Regression with Tfidf Vectorizer = 76.19%
- Extremely Randomized Trees with Count Vectorizer = 74.99%



Conclusion

We were able to construct a model that can predict the nature subreddit a post is in with 76.75% accuracy.

Slightly better at predicting NatureIsMetal correctly (78%) than NatureIsLit (75%)

Model is significantly better than a random prediction, but not reliable for consistently discerning the two nature subreddits